La panne Facebook et BGP

EII-5 / Option ROC - L. Deneire

Six hours with Facebook : network Pb

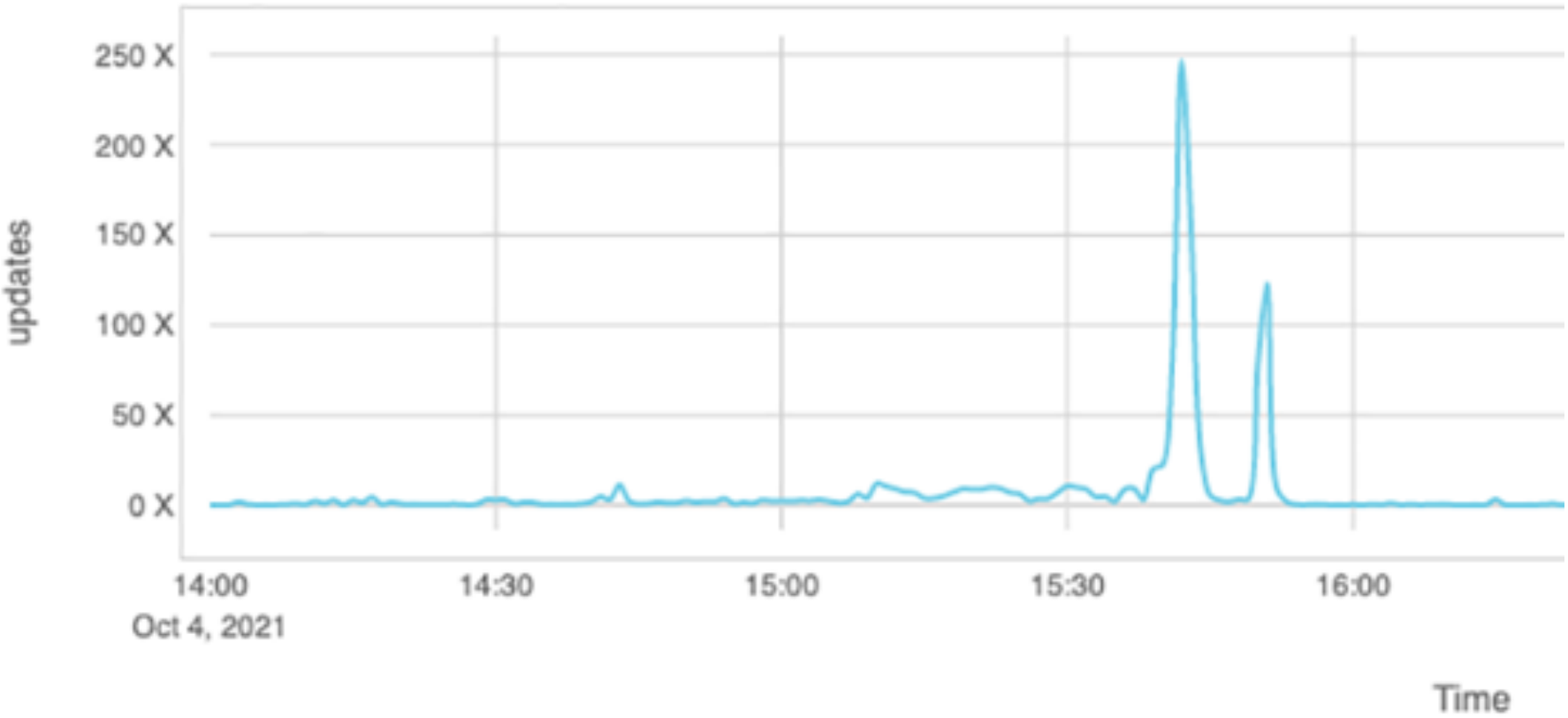
- 6 october 15 :15 UTC, Facebook is down, DNS does not find Facebook ...
- Hence, Whatsapp, Instagram, FB ... Down
- First idea : DNS server down.
- In fact, Facebooks BGP Down

Details in
`https://blog.cloudflare.com/october-2021-facebook-outage`

BGP : routing between big players

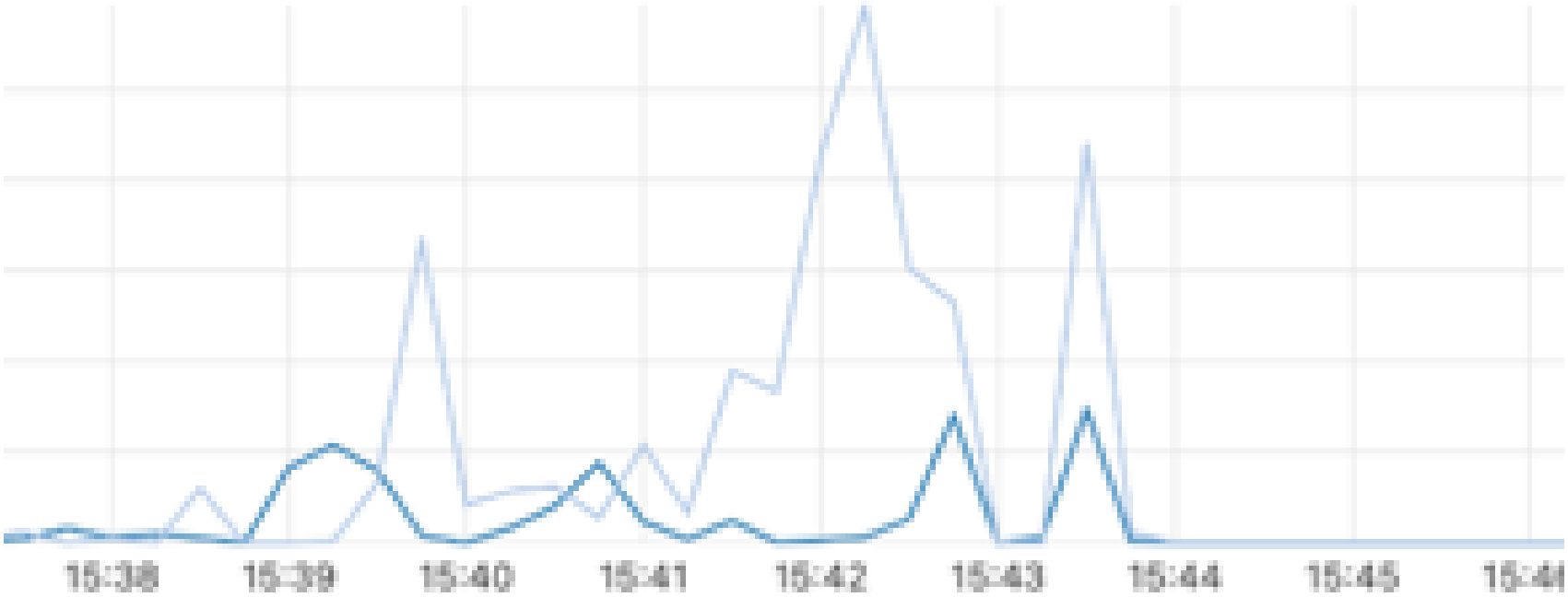
- BGP : Border Gateway Protocol
- Internet : partitionned in Autonomous Systems (AS)
- an Autonomous System is a (potentially big) network : e.g. RENATER.
- ASs have a given number (e.g. RENATER : AS2200)
- have a look at `bgpview.io/reports`
- BGP takes care of routing between ASs

A lot of traffic on BGP around 15 :40 UTC



Details : disconnections

Withdrawals in light blue are predominant (announcements in deep blue)



DNS gets affected

DNS resolvers stopped resolving their domain names

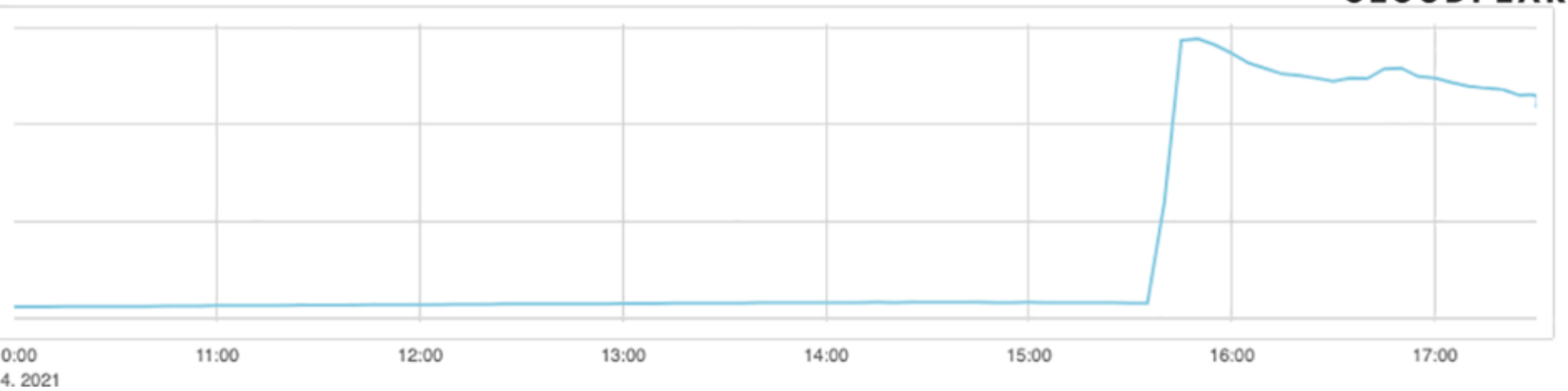
```
➜  ~ dig @1.1.1.1 facebook.com
;; ->>HEADER<<- opcode: QUERY, status: SERVFAIL, id: 31322
;facebook.com.                  IN      A
➜  ~ dig @1.1.1.1 whatsapp.com
;; ->>HEADER<<- opcode: QUERY, status: SERVFAIL, id: 31322
;whatsapp.com.                  IN      A
➜  ~ dig @8.8.8.8 facebook.com
;; ->>HEADER<<- opcode: QUERY, status: SERVFAIL, id: 31322
;facebook.com.                  IN      A
➜  ~ dig @8.8.8.8 whatsapp.com
;; ->>HEADER<<- opcode: QUERY, status: SERVFAIL, id: 31322
;whatsapp.com.                  IN      A
```

Due to Facebook stopping announcing their DNS prefix routes through BGP, other DNS resolvers had no way to connect to their nameservers. Consequently, 1.1.1.1, 8.8.8.8, and other major public DNS resolvers started issuing (and caching) SERVFAIL responses.

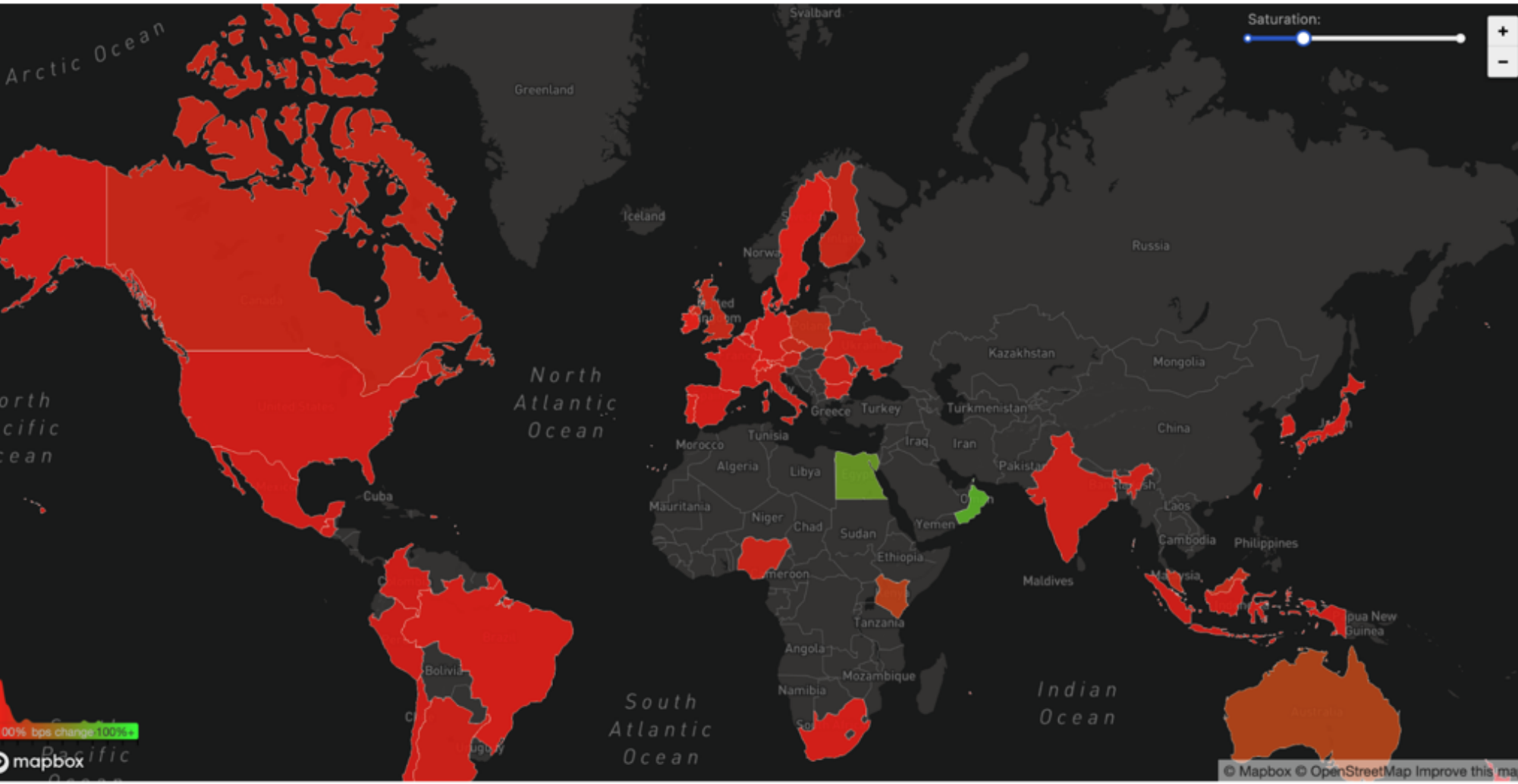
And machines retry, ...

- Applications start retrying, sometimes aggressively
- People click and click again
- And traffic increases on DNS servers (traffic on Cloudflare's DNS)

ueries for websites: facebook, whatsapp, messenger, instagram



And FB traffic decraeses ...



until 21 :00 UTC

And, from 21 :00 UTC on, FB re-announces BGP routes

But who is running BGP

- Large ISPs
- In France : Gandi, RENATER, Free/Proxad, SFR, Orange, OVH, Gitoyen, ...
- They run AS (Autonomous Systems) with ASnumbers (ASN)
- an AS is a network of routers using (mostly) the same policy
- ASNs Free : 12322, Renater : 2200, OVH : 16276, ...

BTW, I set this up fast, from `https://www.bortzmeyer.org/`, Cisco Networking Academy and `https://perso.ens-lyon.fr/eric.fleury/CPS/ART/slides/M1_ART_02-ROUTAGE.pdf`

Peering and Transit

- **Peering** : Connection between Peers (through border routers)
- **Transit** : Connection to a large ISP
- **Tier 1** : ISP that does not by transit (Tata, Levle 3, OpenTransit, ...)
- Internet is basically a result of this peering and transits.

IGP and EGP - Registries

- IGP (Interior Gateway Protocol) : inside of an AS, with a well defined policy, choice of the ISP.
- ECP (Exterior Gateway Protocol) : between ASes, multi-tier, potentially conflicting goal. Obviously only one protocol : BGP.

1. RIRs (Regional Internet Registry) distribute IP prefixes (RIPE-NCC in Europe)
2. LIRs (Local Internet Registry) (some ISPs) are members of a RIR.
3. LIRs ask prefixes to RIRs, RIRs manage the prefix databases.
4. Information through `whois client.rdap.org bfpview.io`

BGP Principle

Border Gateway Protocol, RFC 4271

- Two routers decide to peer
- they establish a TCP connection on port 179 (long lasting connection)
- they announce new routes (`ANNOUNCE`) and outdated routes (`WITHDRAW`)
- BGP only transmits new information (unlike OSPF for example)

BGP Pathes

BGP is using Path vector routing : it uses distance and path information

- A path reads from right to left, the right-most AS is the origin ('4', '5', '1') is a path from 1 to 4 through 5 The BGP prefix PAs (Path Atrributes) are classified :
  1. *Well-known mandatory* : must be included with every prefix advertisement
  2. *Well-known discretionary* : may or may not be included with the prefix advertisement
  3. *Optional transitive :* stays with the route advertisement from AS to AS
  4. *Optional non-transitive :* cannot be shared from AS to AS

  In BGP, the **Network Layer Reachability Information (NLRI)** is the routing update that consists of the network prefix, prefix length, and any BGP PAs for that specific route.

BGP algorithm and loop prevention

- The BGP attribute AS_Path is a well-known mandatory attribute and includes a complete list of all the ASNs that the prefix advertisement has traversed from its source AS.
- AS_Path is used as a loop-prevention mechanism in BGP. If a BGP router receives a prefix advertisement (ANNOUNCE) with its AS listed in AS_Path, it discards the prefix because the router thinks the advertisement forms a loop.

So the algorithm looks like

1. The router refuses announces with including it's own AS, then
2. It takes the one with the best local preference, then if equality
3. It takes the shortest route, then if equality
4. It takes the one with emitted by the router with the smallest ID

No global vision

- Routers have only partial network knowledge
- Hence routers have different information
- Acceptance and propagation are controlled by a local policy (potentially Access Lists - a wrong access list can ruin everything ...)
- Hence BGP is a policy-based routing (rather then technical routing like OSPF)
- and BGP is potentially fragile ... as the Facebook case illustrates

Packet types

Table 11-2 BGP Packet Types

Type	Name	Functional Overview
1	OPEN	Sets up and establishes BGP adjacency
2	UPDATE	Advertises, updates, or withdraws routes
3	NOTIFICATION	Indicates an error condition to a BGP neighbor
4	KEEPALIVE	Ensures that BGP neighbors are still alive

Note that there is a **Hold Time** (typically 180 sec) - if no messages for that duration, the BGP session is torn down and routes are removed.

BGP configuration basics

1. Initialize the BGP process : **router bgp** *as-number*.
2. Configure the BGP router ID (RID) (optional but best practice). **bgp router-id** *router-id* . When the router ID changes, all BGP sessions reset and need to be reestablished.
3. **neighbor** *ip-address* **remote-as** *as-number*
4. Specify the source interface for the BGP session (Optional). **neighbor** *ip-address* **updatesource** *interface-id*
5. Enable BGP authentication (optional) **neighbor** *ip-address* **password** *password* under the neighbor session parameters.
6. Modify the BGP timers (optional). **neighbor** *ip-address* **timers** *keepalive holdtime [minimum-holdtime]*
7. Initialize the address family : **address-family** *afi safi*. Examples of AFIs are IPv4 and IPv6 and examples of SAFIs are unicast and multicast.
8. Activate the address family for the BGP neighbor : **neighbor** *ip-address* **activate**

Simple eBGP topology

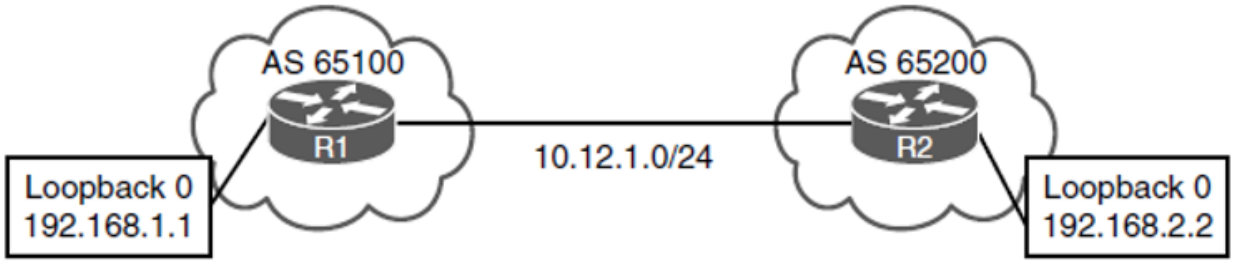


Figure 11-3 *Simple eBGP Topology*

Example 11-2 *BGP Configuration*

```
R1 (Default IPv4 Address-Family Enabled)
router bgp 65100
 neighbor 10.12.1.2 remote-as 65200
 neighbor 10.12.1.2 password CISCOBGP
 neighbor 10.12.1.2 timers 10 40
```

```
R2 (Default IPv4 Address-Family Disabled)
router bgp 65200
 no bgp default ipv4-unicast
 neighbor 10.12.1.1 remote-as 65100
 neighbor 10.12.1.2 password CISCOBGP
 neighbor 10.12.1.1 timers 15 50
 !
 address-family ipv4
  neighbor 10.12.1.1 activate
 exit-address-family
```

BGP Summary

Example 11-3 *BGP IPv4 Session Summary Verification*

```
R1# show bgp ipv4 unicast summary
BGP router identifier 192.168.2.2, local AS number 65200
BGP table version is 1, main routing table version 1

Neighbor        V     AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.12.1.2       4  65200       8       9        1    0    0 00:05:23        0
```

Table 11-3 BGP Summary Fields

Field	Description
Neighbor	IP address of the BGP peer
V	BGP version used by the BGP peer
AS	Autonomous system number of the BGP peer
MsgRcvd	Count of messages received from the BGP peer
MsgSent	Count of messages sent to the BGP peer
TblVer	Last version of the BGP database sent to the peer
InQ	Number of messages queued to be processed from the peer
OutQ	Number of messages queued to be sent to the peer
Up/Down	Length of time the BGP session is established, or the current status if the session is not in an established state
State/PfxRcd	Current state of the BGP peer or the number of prefixes received from the peer

BGP Neighbour

Example 11-4 *BGP IPv4 Neighbor Output*

```
R2# show bgp ipv4 unicast neighbors 10.12.1.1
! Output omitted for brevity

! The first section provides the neighbor's IP address, remote-as, indicates if
! the neighbor is 'internal' or 'external', the neighbor's BGP version, RID,
! session state, and timers.

BGP neighbor is 10.12.1.1, remote AS65100, external link
  BGP version 4, remote router ID 192.168.1.1
  BGP state = Established, up for 00:01:04
  Last read 00:00:10, last write 00:00:09, hold is 40, keepalive is 13 seconds
  Neighbor sessions:
    1 active, is not multisession capable (disabled)
! This second section indicates the capabilities of the BGP neighbor and
! address-families configured on the neighbor.
  Neighbor capabilities:
    Route refresh: advertised and received(new)
    Four-octets ASN Capability: advertised and received
    Address family IPv4 Unicast: advertised and received
    Enhanced Refresh Capability: advertised
    Multisession Capability:
    Stateful switchover support enabled: NO for session 1
  Message statistics:
    InQ depth is 0

! This section provides a list of the BGP packet types that have been received
! or sent to the neighbor router.
                         Sent       Rcvd
    Opens:                  1          1
    Notifications:          0          0
    Updates:                0          0
    Keepalives:             2          2
    Route Refresh:          0          0
    Total:                  4          3
  Default minimum time between advertisement runs is 0 seconds

! This section provides the BGP table version of the IPv4 Unicast address-
! family. The table version is not a 1-to-1 correlation with routes as multiple
! route change can occur during a revision change. Notice the Prefix Activity
```

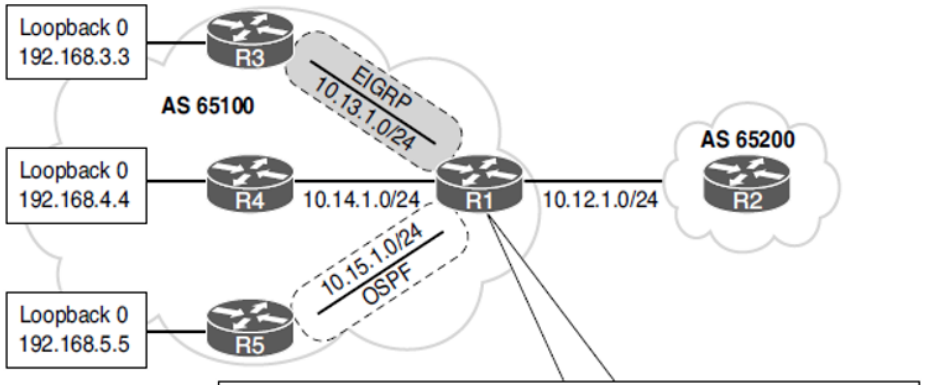
BGP Tables - Prefix Advertisement

BGP uses three tables for maintaining the network paths and path attributes (PAs) for a prefix.

- **Adj-RIB-in** - Contains the Network Layer Reachability Information (NLRI) routes in original form.
- **Loc-RIB** - Contains all the NLRI routes that originated locally or were received from other BGP peers. After NLRI routes pass the validity and next-hop reachability check, the BGP best-path algorithm selects the best NLRI for a specific prefix. The Loc-RIB table is the table used for presenting routes to the IP routing table.
- **Adj-RIB-out** - Contains the NLRI routes after outbound route policies have been processed.

network *network* **mask** *subnet-mask [route-map route-map-name]* : install network prefixes in *Loc-RIB* .
route-map (optional) provides a method to set specific BGP PAs when the prefix installs into the *Loc-RIB* table.

BGP Advertisement




```
R1#show ip route | begin Gateway
Gateway of last resort is not set

    10.0.0.0/8 is variably subnetted, 8 subnets, 2 masks
C       10.12.1.0/24 is directly connected, GigabitEthernet0/0
C       10.13.1.0/24 is directly connected, GigabitEthernet0/1
C       10.14.1.0/24 is directly connected, GigabitEthernet0/2
C       10.15.1.0/24 is directly connected, GigabitEthernet0/3
C       192.168.1.1 is directly connected, Loopback0
D       192.18.3.3 [90/3584] via 10.13.1.3, 00:02:10, GigabitEthernet0/1
S       192.168.4.4 [1/0] via 10.14.1.4
O       192.168.5.5 [110/11] via 10.15.1.5, 00:00:08, GigabitEthernet0/3
```

Figure 11-5 *Multiple BGP Route Sources*

Example 11-5 *Configuration for Advertising Non-Connected Routes*

```
R1
router bgp 65100
 network 10.12.1.10 mask 255.255.255.0
 network 192.168.1.1 mask 255.255.255.255
 network 192.168.3.3 mask 255.255.255.255
 network 192.168.4.4 mask 255.255.255.255
 redistribute ospf 1
```

```
R2
router bgp 65200
 address-family ipv4 unicast
  network 10.12.1.0 mask 255.255.255.0
  network 192.168.2.2 mask 255.255.255.255
```

BGP Database processing

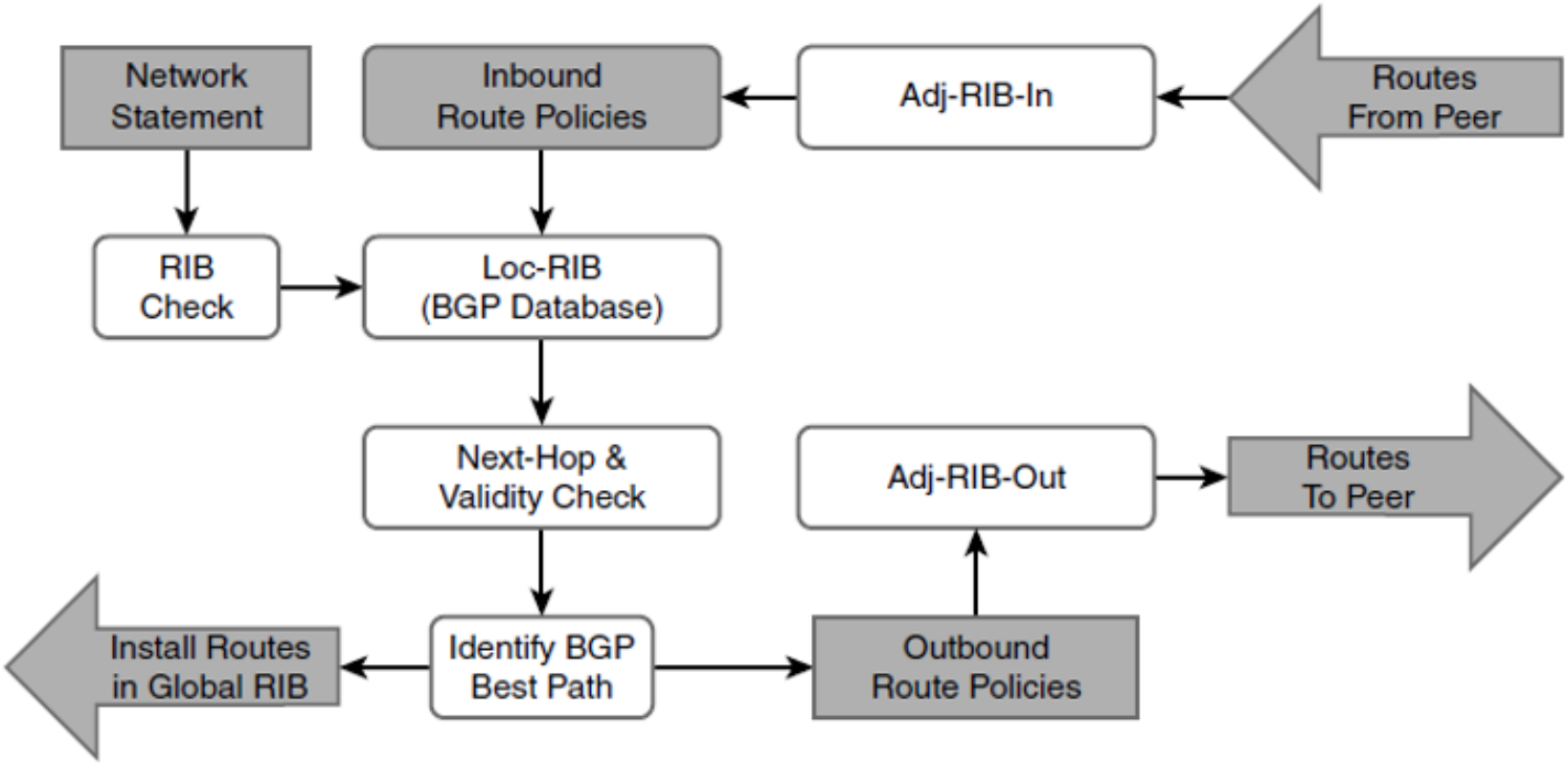


Figure 11-6 *BGP Database Processing*

BGP Attributes

Table 11-5 BGP Prefix Attributes

Output	Description
Paths: (2 available, best #2)	Provides a count of BGP paths in the BGP Loc-RIB table and identifies the path selected as the BGP best path. All the paths and BGP attributes are listed after this.
Advertised to update-groups	Identifies whether the prefix was advertised to a BGP peer. BGP neighbors are consolidated into BGP update groups. If a route is not advertised then *Not advertised to any peer* is displayed.
65200 (1st path) Local (2nd path)	This is the AS_Path for the NLRI as it was received or whether the prefix was locally advertised.
10.1.12.2 from 10.1.12.2 (192.168.2.2)	The first entry lists the IP address of the eBGP edge peer. The 'from' field lists the IP address of the iBGP router that received this route from the eBGP edge peer. (In this case, the route was learned from an eBGP edge peer, so the address is the eBGP edge peer.) Expect this field to change when an external route is learned from an iBGP peer. The number in parentheses is the BGP identifier (RID) for that node.
Origin IGP	Origin is the BGP well-known mandatory attribute that states the mechanism for advertising this route. In this instance, it is an internal route.
metric 0	Displays the optional non-transitive BGP attribute MED, also known as the BGP metric.
localpref 100	Displays the well-known discretionary BGP attribute Local Preference.
valid	Displays the validity of this path.
External (1st path) Local (2nd path)	Displays how the route was learned: internal, external, or local.

BGP details

Example 11-7 *Viewing Explicit BGP Routes and Path Attributes*

```
R1# show bgp ipv4 unicast 10.12.1.0
BGP routing table entry for 10.12.1.0/24, version 2
Paths: (2 available, best #2, table default)
  Advertised to update-groups:
     2
  Refresh Epoch 1
  65200
    10.12.1.2 from 10.12.1.2 (192.168.2.2)
      Origin IGP, metric 0, localpref 100, valid, external
      rx pathid: 0, tx pathid: 0
  Refresh Epoch 1
  Local
    0.0.0.0 from 0.0.0.0 (192.168.1.1)
      Origin IGP, metric 0, localpref 100, weight 32768, valid, sourced, local, best
      rx pathid: 0, tx pathid: 0x0
```

BGP final output table

Example 11-8 *Neighbor-Specific View of the Adj-RIB-OUT Table*

```
R1# show bgp ipv4 unicast neighbors 10.12.1.2 advertised-routes
! Output omitted for brevity
     Network          Next Hop            Metric LocPrf Weight Path
 *> 10.12.1.0/24     0.0.0.0                  0         32768 i
 *>  10.15.1.0/24    0.0.0.0                  0         32768 ?
 *>  192.168.1.1/32  0.0.0.0                  0         32768 i
 *>  192.168.3.3/32  10.13.1.3             3584         32768 i
 *>  192.168.4.4/32  10.14.1.4               0         32768 i
 *>  192.168.5.5/32  10.15.1.5              11         32768 ?

Total number of prefixes 6
```

```
R2# show bgp ipv4 unicast neighbors 10.12.1.1 advertised-routes
! Output omitted for brevity
     Network          Next Hop            Metric LocPrf Weight Path
*> 10.12.1.0/24      0.0.0.0                  0         32768 i
*> 192.168.2.2/32    0.0.0.0                  0         32768 i

Total number of prefixes 2
```

BGP with Ip route

Example 11-9 *BGP Summary with Prefixes*

```
R1# show bgp ipv4 unicast summary
! Output omitted for brevity
Neighbor        V           AS MsgRcvd MsgSent   TblVer  InQ OutQ Up/Down  State/PfxRcd
10.12.1.2       4        65200      11       10        9    0    0 00:04:56            2
```

Example 11-10 *Displaying BGP Routes in an IP Routing Table*

```
R1# show ip route bgp | begin Gateway
Gateway of last resort is not set

      192.168.2.0/32 is subnetted, 1 subnets
B        192.168.2.2 [20/0] via 10.12.1.2, 00:06:12
```